A Review on Multi-modal Speech Representation Learning

Anonymous ACL submission

Abstract

This document is a supplement to the general instructions for *ACL authors. It contains instructions for using the LaTeX style file for ACL 2023. The document itself conforms to its own specifications, and is, therefore, an example of what your manuscript should look like. These instructions should be used both for papers submitted for review and for final versions of accepted papers.

1 Introduction

The goal of learning intermediate latent representations of a given input instead of learning an explicit task end-to-end is to be able to extract information out of the input and use this information to tackle multiple tasks. Speech representations are latent representation of speech input. This input could be in formats such as raw waveform, or mel filterbank features. Speech representations are related to (textual) semantic word embeddings as these embeddings are also latent representations with extracted information, in that case their semantic relatedness. (Pennington et al., 2014) The difference from textual embeddings to speech representations and what makes learning speech representations hard is that spoken words don't have clear boundaries, speech input is continuous as opposed to discrete textual words and that speech contains more information than text like for example speaker information, noise or emotion.

The first attempts at learning speech representations were done using clustering algorithms like k-means or Gaussian Mixture Models (Rabiner and Wilpon, 1979) and later improved by adding Hidden Markov Models to allow processing of continuous speech rather than single words. (Bahl et al., 1986) Currently, the prevalent approach for learning speech representations is to perform pretext task optimization, also known as pre-training. In this approach, the representations are learned by solving a task that is derived by unlabeled data like for example predicting the next word in a sequence of words or predicting a masked word in a sequence. The advantage of this approach is that it only needs unlabeled data which is of higher availability than labeled data. Within the pretext task optimization approach, the following three learning paradigms can be considered the most widely used:

1. *Generative learning*: In generative learning, an input is reconstructed based on a limited view of it. This can mean predicting the next word in a sequence, a masked word within an input or predicting the original from a noisy input (Vincent et al., 2008).

2. *Contrastive learning*: Contrastive learning is based on the idea of...

3. *Predictive learning*: Predictive learning is a...

This paper is structured as follows. First, single-mode speech representation models are introduced, specifically wav2wec2.0 and HuBERT which is the basis of some of the multi-modal speech representation models. Afterwards, multi-modal speech representation models are shown by first looking at their architecture, then the learning approach and specifics and finally the performance of the models is discussed with respect to other multi-modal and also single-mode models. Finally, a discussion about the impact, advantages and challenges of multi-modal speech representation learning is conducted including a comparison to single mode approaches.

1.1 History of Speech Representation Learning

Show first approaches (?)

1.2 Self-Supervised Learning for Speech Representation Models

1. Motivation for self supervised learning / What is self supervised learning in general. Where does it maybe originate from?
With the rise of deep learning, the use of labeled data to train models capable of . . . HOW TO INTRO? WITHOUT USING WORDS OF (Mohamed et al., 2022)?

One example of labeled speech data is paired audio-text data, which consists of pairs of voice tracks and corresponding text segments. These kind of data can for example be used to train end-to-end automatic speech representation (ASR) models.
Human-labeled speech data is expensive and generally of limited supply, especially so for languages with comparatively few speakers in the world.
This is why methods using only unlabeled speech data were developed and are since being used to tackle many natural language processing tasks such as the aforementioned ASR task (Kemp and Waibel, 1970).

Self-supervised learning comprises of "techniques that utilize information extracted from the input data itself as the label to learn representations useful for downstream tasks." (Mohamed et al., 2022)
2. Bridge to speech/natural language processing
3. Pre-training. Generative Learning. Contrastive Learning. Predictive Learning.

2 Single-mode Speech Representation Models

2.1 wav2vec2.0

2.2 HuBERT

3 Multi-modal Speech Representation Models

This section presents three examples of speech representation learning models each leveraging a different set of input channels. First, we will look at SpeechT5 (Ao et al., 2022) which is using the additional text modality for learning speech representations. Next, we will see AV-HuBERT (Shi et al., 2022), which is an extension of the single-mode model HuBERT () and benefits from the audio modality as well as the video modality. Lastly, VAT-LM () will be covered, leveraging all three mentioned modalities: audio, video and text.

Model	WER (%)
wav2vec2.0	1.8
SpeechT5	1.5

Table 1: ASR Performance.

3.1 SpeechT5

The SpeechT5 framework, first introduced in 2021 by Microsoft (Ao et al., 2022), is an expansion of the text-only T5 framework (Text-to-Text Transfer Transformer, Raffel et al.'s (2023)). SpeechT5 is. . .

3.1.1 Model Architecture

3.1.2 Learning (?)

3.1.3 Performance Discussion

Here, we show that SpeechT5 outperforms single-mode SRL models, and also the benefit of using multiple modalities by presenting the ablation studies of the original paper. See Table 1. Also, here is a section of the appendix containing more information: A.

3.2 AV-HuBERT

AV-HuBERT Stuffs

3.2.1 Model Architecture

3.2.2 Learning (?)

3.2.3 Performance Discussion

3.3 VAT-LM

VAT-LM Stuffs

3.3.1 Model Architecture

3.3.2 Learning (?)

3.3.3 Performance Discussion

4 Discussion

Discussion on how well multi-modal models improve the field of SRL.

Ethics Statement

Scientific work published at ACL 2023 must comply with the ACL Ethics Policy.[1] We encourage all authors to include an explicit ethics statement on the broader impact of the work, or other ethical considerations after the conclusion but before the references. The ethics statement will not count toward the page limit (8 pages for long, 4 pages for short papers).

[1] https://www.aclweb.org/portal/content/acl-code-ethics

Acknowledgements

Maybe acknowledge the review (Mohamed et al., 2022) or smth.

References

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022. SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing. ArXiv:2110.07205 [cs, eess].

L. Bahl, P. Brown, P. de Souza, and R. Mercer. 1986. Maximum mutual information estimation of hidden markov model parameters for speech recognition. In *ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 11, pages 49–52.

Thomas Kemp and Alex Waibel. 1970. Unsupervised training of a speech recognizer: Recent experiments. *Proc Eurospeech Budapest*, 6.

Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D. Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, Tara N. Sainath, and Shinji Watanabe. 2022. Self-Supervised Speech Representation Learning: A Review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210. ArXiv:2205.10643 [cs, eess].

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

L. Rabiner and J. Wilpon. 1979. Considerations in applying clustering techniques to speaker independent word recognition. In *ICASSP '79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 578–581.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. ArXiv:1910.10683 [cs, stat].

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction. ArXiv:2201.02184 [cs, eess].

Pascal Vincent, Hugo Larochelle, Y. Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. pages 1096–1103.

A Example Appendix

This is a section in the appendix.